Evaluating an Explicit Instruction Teacher Observation Protocol through a Validity Argument Approach

Evelyn S. Johnson, Yuzhu Zheng, Angela R. Crawford, and Laura A. Moylan

Boise State University

Author Note

Evelyn S. Johnson, Ed.D., Department of Early and Special Education, Boise State University; Yuzhu Zheng, Ph.D., Project RESET; Boise State University, Angela Crawford, Ed.D., Project RESET, Boise State University; Laura A. Moylan, MEd., Project RESET, Boise State University.

This research was supported the Institute of Education Sciences, award number R324A150152 to Boise State University. The opinions expressed are solely those of the authors. Correspondence regarding this manuscript should be addressed to: Dr. Evelyn S. Johnson, Boise State University, 1910 University Dr., MS 1725, Boise, Idaho 83725-1725. Email: evelynjohnson@boisestate.edu

Citation: Johnson, E. S., Zheng, Y., Crawford, A. R., & Moylan, L. A. (2020, in press). Evaluating an Explicit Instruction Teacher Observation Protocol through a Validity Argument Approach. *Journal of Experimental Education.*

Abstract

In this study, we examined the scoring and generalizability assumptions of an Explicit Instruction (EI) special education teacher observation protocol using many-faceted Rasch measurement (MFRM). Video observations of classroom instruction from 48 special education teachers across four states were collected. External raters (n = 20) were trained to observe and evaluate instruction using the protocol. The results of this study suggest that the scoring rule is appropriate, in that the three-point scale allows for a meaningful way to differentiate various levels of quality of implementation of EI across teachers. Raters consistently scored easier items with higher scores, and more difficult items with a lower score. Additionally, the MFRM results for the rater facet suggest that raters can consistently apply the scoring criteria, and that there is limited rater bias impacting the scores. Implications for research and practice are discussed.

Keywords:

Rater accuracy; teacher observation; rater consistency; feedback; special education

Evaluating an Explicit Instruction Teacher Observation Protocol through a Validity Argument Approach

A long-standing assumption in the field of special education has been that as teachers learn more about the practices that have been found to be effective in improving student outcomes, they would use these in their classroom (Odom, et al., 2020). This assumption however, has proven to be largely false. For example, despite the decades of research supporting explicit instruction as an evidence-based practice (EBP) for students with or at-risk for high-incidence disabilities (SWD; Hughes et al. 2017; Stockard et al., 2018), observation studies of special education instructional practice suggest that explicit instruction has not been effectively implemented on a large scale (Ciullo et al., 2016; McKenna et al., 2015; Swanson, 2008). Explicit instruction was recently identified as one of 22 high leverage practices (HLPs) for students with disabilities by the Council for Exceptional Children (McLeskey et al., 2017). Explicit instruction is an instructional approach that is highly effective for students with high incidence disabilities (SWD) and is supported by nearly 50 years of research (Hughes et al., 2017; Stockard et al., 2018). This research to practice gap may explain why the significant achievement gap between SWD and their peers without disabilities persists (Gilmour, et al., 2019; Schulte & Stevens, 2015; Wei, et al., 2011).

The potential benefit of EBPs to improve outcomes for students are limited by the lack of sustained, quality implementation (Cook & Odom, 2013), and therefore, the challenge remains: *how* to achieve and sustain fidelity of implementation of practices in the classroom once they have been established as effective in the research.

Fidelity of implementation is often described and defined in the literature in terms of two broad dimensions: structural and process (Harn et al., 2013; Odom, 2009; O'Donnell, 2008).

Structural dimensions include things such as adherence to all of the steps of an intervention, or the duration and frequency of implementation, whereas process dimensions focus on the *quality* of delivery of the intervention (Harn et al., 2013; Justice et al., 2008). Quality of delivery has been shown to impact the effect of EBPs on desired outcomes, and is arguably the most important, yet also the most difficult aspect of fidelity to achieve (Harn et al., 2013; Mowbray et al., 2003). In research settings, quality of implementation measures are typically developed by researchers, who devise checklists that include the steps of a practice, often rated on a yes/no scale, with a small percentage of observations evaluated by two or more raters to establish some indication of interrater reliability. Once a practice is established as effective in the research, there is usually no attempt to examine whether the scoring criteria that were developed for research purposes can be appropriately applied within a practice setting to inform meaningful distinctions in the levels of quality of implementation. Additionally, there is rarely an opportunity to determine whether practitioners who are charged with overseeing the adoption of a new practice are able to effectively and reliably use these measures to observe, evaluate and provide feedback to teachers. This limited approach to developing fidelity assessment instruments does little to provide the fidelity assessment tools that the implementation science research has indicated as necessary to promote stronger, sustained implementation of EBPs in the classroom (National Implementation Research Network [NIRN], 2020).

Fidelity Assessment

Three main approaches to improving the fidelity of implementation of evidence-based instructional practices have been described in the literature: (a) creating detailed descriptions of the specific elements of an intervention for practitioners to implement; (b) developing reliable measures of the quality of intervention of these elements; and (c) providing ongoing observation

and feedback aligned with the use of these measures (Carroll et al., 2007; Hill & Grossman, 2013). Validated teacher observation systems developed in this way can promote quality implementation by identifying and defining effective practice, serving as a guide to teachers, providing opportunities for feedback and informing professional development needs (Hill & Grossman, 2013; Papay, 2012). Emerging evidence supports the effectiveness of some observation systems for improving instruction and students' academic outcomes (Dee & Wyckoff, 2015; Steinberg & Sartain, 2015; Taylor & Tyler, 2012). However, few existing observation instruments are developed at a level of specificity to support the implementation of specific, evidence-based practices and fewer still are rigorously tested to ensure they result in psychometrically sound measures of the quality with which EBPs are implemented (Hill & Grossman, 2013). There is a growing consensus about the need for precise and reliable measures of the critical elements of instructional practice across a range of content areas to promote the consistent use of evidence-based practices in the classroom (Munter, 2014; O'Donnell, 2008; Sullivan et al., 2016).

RESET Explicit Instruction Observation Protocol

One effort to create a set of psychometrically sound teacher observation instruments is the Recognizing Effective Special Education Teachers (RESET) observation system, a federally funded project to align observation protocols with EBPs for students with or at-risk for high-incidence disabilities (SWD). The goal of RESET is to leverage the extensive research on EBPs for this population of students and to develop observation protocols aligned with these EBPs in order to: (a) determine the extent to which teachers are implementing EBPs with fidelity, (b) provide feedback to teachers to improve their practice and ultimately, (c) improve outcomes for SWD (Authors, 2018). The theory of action underlying RESET rests on the idea that improving

teacher practice depends upon establishing a clear target for quality instruction aligned with the salient characteristics of EBPs, then orienting feedback and measuring improvement with the observation protocol of interest. Through this process, it is anticipated that the teachers' quality of implementation of EBPs improves, and this instructional improvement results in accelerated student growth. RESET was developed using the principles of Evidence-Centered Design (ECD; Mislevy et al., 2003) and consists of 21 rubrics that detail instructional practices organized in three categories, 1) instructional methods, 2) content organization and delivery, and 3) individualization. A complete description of how the RESET rubrics were developed is provided elsewhere (see Johnson et al., 2018). In this study, the focus is on the Explicit Instruction rubric, which has been designed to evaluate teachers' quality of explicit instruction implementation. The RESET EI observation protocol is comprised of 25 items that detail the elements of EI as described in the research, across three performance descriptor levels (see Figure 1 for a list of items for the 'implemented' descriptor). To use RESET, teachers submit video recordings of their lessons, which are then observed by trained raters. The teachers' level of implementation of each item is evaluated on a three-point scale in which a score of 3 is implemented, a 2 is partially implemented, and a 1 is not implemented. Raters are trained through the use of exemplars and elaborated descriptions and examples of practice at each of the three levels of performance. Raters then view recorded lessons between 20-45 minutes in length, and assign a score for each item on the rubric, citing the evidence they used within the observed lesson to reach their scoring decision. Both item scores and an overall lesson score are reported to the teacher. A lesson score is based on the average performance across the items to provide an overall assessment of a teacher's quality of implementation of Explicit Instruction. Given the intended use of RESET, item-level scores are important, as they provide the teacher with specific, actionable feedback

about which elements of Explicit Instruction they are implementing well, and which elements they may need to improve.

To support these uses of RESET, several assumptions about the resulting scores and their generalizability need to be met (Authors, 2020). In Table 1, we present the validity argument (Kane, 2006) and assumptions for RESET that have guided the research agenda to establish a psychometrically sound instrument. In the current stage of development, our focus has been on testing the scoring and generalizability assumptions. Assumptions about scoring include ensuring that the scoring rules are appropriate (e.g., the performance descriptors and associated scores result in meaningful distinctions with regard to quality of implementation), that raters' can consistently apply the scoring criteria and can use the items without bias in that the same instructional behaviors and quality observed across different teachers will be scored similarly (Bell et al., 2012; Kane, 2006). Assumptions about generalizability include that teacher performance is generalizable across raters and lessons (Bell et al., 2012; Kane, 2006). As mentioned, observation systems hold significant promise as a means to improve the fidelity with which evidence-based instructional practices are implemented, but without evidence to support these scoring and generalization assumptions, it would be premature to use RESET to make decisions (e.g. give feedback or provide an evaluation score) about teacher performance. Although the EI protocol has been found to result in reliable evaluations of teacher practice across several, small studies (Authors, 2018; Authors, 2018; Authors, 2019), the studies to date have been conducted with relatively small samples. Given the long-term goal to establish an observation system that can be used to reliably measure and improve the fidelity with which Explicit Instruction is implemented in the classroom, the purpose of the current replication study was to examine the evidence to support the scoring and generalization assumptions of the

RESET EI observation protocol with a larger sample of teachers and raters. Specifically, the research questions that guided this study were:

1. Is the scoring rule for the RESET EI protocol appropriate to identify meaningful distinctions in the quality of explicit instruction implementation?
2. Can raters consistently apply the RESET EI protocol scoring criteria?
3. Does the RESET EI protocol result in reliable estimates of the quality of a teacher's implementation of explicit instruction?

Method

Participants

Special education teachers. Forty-eight special education teachers from 34 schools, and 17 districts across 3 states participated in this study. Table 2 provides detailed information of the teachers' demographics. Teachers were recruited by sending study information and recruitment letters to the special education district directors, who shared recruiting materials with their special education teaching staff. All participants were female, teaching from kindergarten to 10th grade levels in an intervention setting. Teachers ranged in age, with 10 teachers who were 20-29 years old, 14 teachers between 30-39 years old, 11 teachers between 40-49 years, nine teachers between 50-59 years old, and two were 60 or over 60 years old. Two teachers declined to provide information on their age. One teacher was Asian, two were Hispanic, and the remaining teachers were white. Their number of years' experience ranged from 1 to 36 years (M=10.83, SD=9.15). Thirty-one participants held a Bachelor's degree and 15 teachers held a Master degree in education or special education. Two teachers declined to give information about their level of education. Thirty-three teachers taught reading, 12 taught math, and three teachers taught both math and reading. When asked to report what instructional practice they used, three teachers

declined to give information. Among the rest of the teachers, four reported using a curricular program based on explicit instruction practices. Twenty teachers reported using direction instruction. Fourteen reported using a curricular program based on direct instruction and explicit instruction. Seven teachers reported using direct instruction, explicit instruction and cognitive strategy instruction.

Raters. A total of 21 raters including 16 females and five males were recruited from seven states in this study. Raters were recruited and selected based on the following criteria: (a) holding a teacher certificate, (b) having three or more years of experience in special education or closely related field, (c) have experience with teacher observation and (d) training and/or experience delivering explicit instruction to SWD. One rater dropped out of the study after the rater training. Among the remaining 20 raters, 17 were white, two were Asian-American, and one was Turkish. Raters had between 3 to 20 years of working experience in special education. Thirteen raters had a Master's Degree, six had a Doctoral Degree, and one rater had a Bachelor's degree. At the time of the study, seven raters worked as classroom special education teachers, six were in doctoral degree programs, five worked as a special education faculty or researcher at a university, one was a State Education Specialist, and one worked as an education curriculum developer. Eleven raters reported having completed formal coursework in Explicit Instruction when they were in the undergraduate or graduate program, two raters were trained in Explicit Instruction supervision during graduate school, and seven raters reported learning Explicit Instruction as in-service teachers.

RESET Explicit Instruction Observation Protocol

The RESET *Explicit Instruction* observation protocol (see Appendix A) was used to evaluate participating teachers' quality of explicit instruction implementation. This observation

protocol consists of 25 items that detail the elements of explicit instruction (see Authors et al., 2018 for a description of the RESET observation system development process). Each item is rated on a three-point scale (1=Not implemented, 2= Partially implemented, and 3 =Implemented) to evaluate a teacher's level of proficiency in implementing that specific element. Studies using small samples of teachers and raters suggest that the Explicit Instruction protocol provides reliable assessments of a teacher's quality of implementation of this EBP (Authors et al., 2018; Authors et al., 2018; Authors et al., 2019).

Procedures

Video collection. All teachers provided video-recorded lessons of their instruction during at least one of the 2015-16 through 2018-19 school years. Videos were recorded and uploaded using the Swivl™ capture system and ranged in length from 20 to 60 minutes. Swivl™ is a movable bot that accommodates a video recording device (e.g., tablet or smart phone) that is uploaded through an app that allows for the management of video recorded instruction. Each teacher contributed a total of 20 videos, for a total of 960 videos. For the purpose of this study, three videos from the same school year were selected from each teacher. Videos had to have adequate video and audio quality and had to depict a lesson which reflected their explicit instruction practice. A total of 144 videos were selected to be evaluated in this study. All selected videos were then assigned an identifying number, and listed in random order to control for order effects.

Rater training and scoring. Over a four-day training period, raters were provided with an overview of the RESET project goals and a description of how the EI protocol was developed. Research project staff then explained each item of the EI protocol and clarified any questions the raters had. Raters were provided with a training manual that includes detailed descriptions of

each item, along with examples for each item across each level of performance. Then, raters watched and scored a video that had been scored by project staff. The scores were reviewed and discussed to include the rationale for the score that each item received. Raters then watched and scored three videos independently, and scores were reconciled with the master scored protocol. Disagreements in scores were reviewed and discussed.

Raters were then assigned a randomly ordered list of videos and asked to evaluate the videos in the assigned order, to score each item, to provide time stamped evidence used as a basis for the score, and to provide a brief explanation of the rationale for their score. Raters were reminded to consult the training manual as they completed their observations and were given a timeframe of six weeks to complete their ratings. Completed evaluations were submitted using an electronic version of the protocol developed in the Qualtrics ® survey system. To maintain a feasible observation load, we developed a rating scheme that allowed for scores across raters and videos to be linked without requiring each rater to score each video (Eckes, 2011). We randomly selected two teachers to have their first and last video scored by every rater. One rater was randomly selected to score at least one video of each teacher. Remaining videos were randomly assigned and each video was scored by three raters, and each teacher was scored by either nine or ten raters. This created a design in which all raters scored 25 videos.

Data Analysis

Many-faceted Rasch measurement (MFRM) analyses were conducted to analyze the scores assigned to the recorded lessons by raters. MFRM is a model including all sources of variability (facets) that are thought to influence the scores in the analysis (Eckes, 2011). All facets are calibrated simultaneously and receive a common score on a linear scale (the logit scale) that represents the latent construct. Each facet can be examined independently to assess

levels of reliability, precision, and consistency to help determine whether or not the rating system is functioning as intended (Vogel & Engelhard, 2011). All facets can be examined on a common measurement level through the calibration map allowing for all facets to be interpreted simultaneously (Linacre, 2014). This study was designed as a four-facet model including item, teacher, rater, and lesson. The model used for the MFRM analysis in this study is given by:

$$\ln P_{nijok} P_{nijo(k-1)} = B_n - D_i - C_j - T_o - F_k$$

where P_{nijok} is the probability of teacher *n*, when rated on item *i* by judge (rater) *j* on occasion (lesson) *o*, being awarded a rating of *k*. $P_{nijo(k-1)}$ is the probability of teacher *n*, when rated on item *i* by judge *j* in occasion *o*, being awarded a rating of *k*-1, B_n is the ability of teacher *n*, D_i is the difficulty of item *i*, C_j is the severity of judge *j*, T_o is the stringency of occasion *o*, and F_k is the difficulty overcome in being observed at the rating *k* relative to the rating *k*-1 (Eckes, 2011).

The MFRM analysis was conducted by means of the computer program FACETS (version 3.81; Linacre, 2019). This program used the ratings that raters awarded to teachers to estimate individual teacher proficiency, rater severities, item difficulties, and lesson difficulties. MFRM analysis produces infit and outfit statistics for each facet, two quality control statistics that indicate whether the measures have been confounded by construct-irrelevant factors (Eckes, 2011). Ranges in fit statistics from .5 to 1.5 are considered acceptable (Eckes, 2011; Englehard, 1992). FACETS also provides reliability and separation indices. The reliability index indicates the reproducibility of the measures if the test were to be administered to another randomly selected sample from the same population (Bond & Fox, 2007). Separation indicates the number of statistically distinguishable strata in the data. Additionally, for the teacher facet, FACET also provides fair scores or fair averages, which result from a transformation of teachers' proficiency estimates reported in logits to the corresponding scores on the raw-scale score (Eckes, 2011). A

fair average is the score that a particular teacher would have obtained from a rater of average severity and illustrates the effect of the model-based compensation for differences in rater severity (Eckes, 2011). A full analysis of teacher, rater, lesson and item analyses is reported in this study.

Results

MFRM analyses were used to examine the statistics for the four facets (item, teacher, rater and lesson) in this study, summarized in Figure 2 and Tables 2-5. Exact rater agreement was 52%. All analyses are based on a total of 12,343 assigned scores. Category statistics shows that of the assigned scores 45% were a 3 (implemented), 42% were a 2 (partially implemented), and 13% were a 1 (not implemented).

Item Facet and Fit Statistics

Figure 2 is a variable map presenting vertical rulers that display the location of elements on the logit scale for each facet to allow for the comparison within and among the facets. The scale along the left, titled "Measr," is the logit measure for the elements within each facet. The scale, estimated from the pattern of the data, ranged from -2 to +4. A higher location on the vertical ruler indicates more difficulty for items, higher proficiency for teachers, more severity for raters, and more difficulty for lessons. As shown in Figure 2, there is a considerable range between the most difficult item (or more accurately, the item that few teachers implemented well) which was Item 3, *The teacher clearly explains the relevance of the stated goal to the students* was the most difficult, and the lease difficult item, Item 19, *The teacher provides frequent opportunities for students to engage or respond during the lesson*.

Table 3 reports the item difficulty, fit statistics, separation and reliability indices for the item facet. As depicted in Figure 2, and reported in more detail in Table 3, the item difficulty

ranges from 2.24 logits (*SE*=.08) for Item 3, to -1.10 logits (*SE*=.09) for Item 19. The fit statistics ranged from .78 to 1.63 (Item 3), which is slightly higher than the upper bound of the acceptable range of .50 to 1.50 (Eckes, 2011). This suggests that raters consistently scored easier items with higher scores, and more difficult items with a lower score, but that this pattern may have been broken for Item 3. Outfit statistics are sensitive to extreme values (Engelhard, 1994), and in the present analysis the higher outfit statistic for Item 3 is likely the result of teachers who otherwise performed well on the other items of the protocol receiving a low score for this item. The item reliability of separation of .99 demonstrates that item difficulties are separated along the continuum of difficulty of explicit instruction implementation. This separation was statistically significant with a chi-square of 1703.3, 24 degrees of freedom and significance < .001.

Teacher Facet and Fit Statistics

The teacher column on Figure 2 lists the teachers from most proficient (Teacher 11 and Teacher 45) to least proficient (Teacher 24) at the bottom. Teachers who are more proficient are expected to score higher than teachers who are less proficient on items that are more difficult. Table 4 reports the teachers' overall fair average score on the EI protocol, the fit statistics, and provides the reliability and separation indices for the teacher facet. The teachers' performance on the EI protocol ranges from 3.05 logits (*SE*=.20) for Teacher 11, who is the most proficient, to -.94 logits (*SE*=.12*)* for Teacher 24, who is the least proficient. The fair average score ranges from 1.60 for Teacher 24 to 2.89 for Teacher 11. The fit statistics measure the extent to which a teacher's pattern of responses matches that predicted by the model, and therefore can be used to identify teachers who have been evaluated in a consistent manner. Table 4 shows that all fit statistics fell between .66 to 1.49, which are within acceptable levels (Eckes, 2011), suggesting

that the evaluation with the protocol has been consistently applied to determine teachers' ability to implement explicit instruction. The reliability of separation is .98, with a statistically significant chi square of 1905 and 47 degrees of freedom ($p<.001$). This indicates that teachers differ in the quality of their implementation of explicit instruction as measured by this protocol, beyond what can be attributed to measurement error.

Rater Facet and Fit Statistics

The rater column on Figure 2 ranks the raters from most severe (Rater 17) at the top to the most lenient (Raters 7 and 9) at the bottom. Table 5 shows that the raters' severity ranges from .62 logits (*SE*=.07) to -.67 logits (*SE*=.07). The fit statistics help determine whether raters are consistent with their own ratings on the protocol and can be used to identify severe or lenient ratings that are unexpected given a rater's overall scoring pattern, or used to identify biases for a particular item or teacher. Fit values greater than 1 show more variation than expected (misfit), and values less than one show less variation than expected (overfit). Misfit is generally thought to be more problematic than overfit (Myford & Wolfe, 2003). The fit statistics fell within .74 to 1.27, which are within acceptable levels (Eckes, 2011). The reliability coefficient was .96, on a chi-square of 518.3 and 19 degrees of freedom ($p < .001$) and separation was 5.17, which demonstrate reliable difference in rater's severity.

Lesson Facet and Fit Statistics

Of the four facets, the Lesson facet shows the least variability, as shown in Figure 2. The lesson facet is also somewhat difficult to interpret because we did not specify the content or focus of the lessons other than the teacher had to use explicit instruction as the primary instructional method. Table 6 shows that the lessons' difficulty ranges from .16 logits (*SE*=.03) for Lesson 3 which is the most difficult, to -.14 logits (*SE*=.03) for Lesson 1 which is the easiest.

The fit statistics fell within .96 to 1.03, which are within acceptable levels (Eckes, 2011). The reliability of separation of .97 was significant ($p<.001$).

Discussion

The results of our analyses are consistent with those reported in our previous research, which together suggest the RESET EI observation protocol can be used to provide consistent evaluations of the quality with which EI is implemented in the classroom. This is an important finding, because EI has been shown through numerous studies to have a positive effect on student outcomes. Additionally, EI was recently identified as a high-leverage practice (HLP) by the Council for Exceptional Children (McLeskey et al., 2019). When considered along with reports across observation studies that EI is not widely used in practice, the need for fidelity assessments that facilitate improving practitioners' quality of EI implementation are needed if SWD are to reap the benefits of this EBP.

Measures of fidelity to an EBP are often central components of a comprehensive implementation science plan (Bauer et al., 2015). In order for fidelity measures to inform continuous improvement, it is critical that the protocols used are psychometrically sound. In the current study, we collected evidence to examine several assumptions of the scoring and generalizability assumptions that guide the development and research agenda around RESET. The results of this study, along with our previous findings, suggest that the scoring rule is appropriate, in that the three point scale allows for a meaningful way to differentiate various levels of quality of implementation of EI across teachers. Raters consistently scored easier items with higher scores, and more difficult items with a lower score. Additionally, the MFRM results for the rater facet suggest that raters can consistently apply the scoring criteria, and that there is limited rater bias impacting the scores.

With evidence to support the scoring and generalizability assumptions, the RESET EI observation protocol has the potential to meaningfully address the research to practice gap in that it provides detailed descriptions of the specific elements of an intervention for practitioners to implement (Carroll, et al., 2007; Hill & Grossman, 2013); and also provides reliable measures of the quality of implementation of these elements. These are important first steps in the long, arduous process of validating fidelity assessments that can enhance the wide-scale implementation of effective practices, but it can take many years to conduct the studies needed to validate a measure of fidelity by demonstrating that its use correlates with positive outcomes (NIRN, 2020). Future studies are needed to examine the extrapolation and decision assumptions of RESET.

Although the RESET EI protocol shows strong promise in our studies thus far, additional research will be needed to validate this system for use in ***practice.*** Our studies of RESET to date have been conducted within a *research context*, in which multiple raters evaluated multiple lessons provided by participating special education teachers in a low-stakes environment. Although our findings indicate that we can obtain consistent measures of special education teachers' instruction under controlled, experimental conditions, the observation system cannot currently be considered reliable and valid when applied across different settings (e.g. when used in schools by school-based personnel). There are critical differences in the way in which observation systems are implemented in practice that affect their validity (Liu, Bell, Jones & McCaffrey, 2019), and these differences will need to be understood if RESET is to fulfill its potential as a way to improve fidelity of implementation of EBPs for SWD.

For example, in an analysis of the FFT used in "practice-based contexts" as compared to a "research-based context" Liu et al. (2019) found considerable variability in scoring

distributions, scoring tendencies, and in various aspects of the observation system, including the number of lessons observed, and the number of raters observing each lesson. Each of these aspects has been found to significantly impact the reliability and validity of teacher observation results. Their findings led them to recommend, as others have, that observation ***systems*** as opposed to just observation ***protocols*** should be the unit of analysis for validation efforts, and that observation systems validated in research cannot be assumed to be valid for use in practice (Hill et al., 2012; Liu et al., 2019).

Limitations

Although the results of this study are promising, there are several limitations that warrant caution when generalizing results. First, the heterogeneity of our teacher sample (e.g. all female, predominantly White), limits the generalizability of the study's findings. Second, teachers' ability to implement EI was based on three lessons sampled from different time points across the school year. Research in teacher observation has shown that teacher performance can vary depending on the time of year and based on the students in class (Mantzicopoulos et al., 2018). However, several studies have indicated that between 2-4 observations of teachers' instruction can provide reliable estimates of teacher performance (Crawford et al., 2018; Kane & Staiger, 2012). Our findings found little variability in teacher performance as a result of the lesson, suggesting that this may not have been an issue in the current data set. A third limitation is the variability in the length of the videos reviewed. The lessons included in this study ranged from 20 to 60 minutes, which represents a broad range across lesson time and introduces the concern of rater fatigue. To mitigate this concern, the assigned videos for each rater were varied in length (e.g., all raters scored some shorter and some longer videos), and the MFRM fit statistics and bias analyses did not indicate any consistent differences in scoring as a result of the lesson

observed, suggesting that the length of the video may not have impacted raters' ability to provide consistent ratings across lessons.

Finally, the RESET EI observation protocol focuses on one instructional EBP for students with high-incidence disabilities. The EI protocol does not include items related to classroom management or other important dimensions of teaching. These dimensions of teaching are captured by other teacher observation systems, and an additional concern about implementing observation systems in practice is the need for practitioners to be well versed in the use of these multiple systems.

The effectiveness research on explicit instruction suggests an urgent need to support special education teachers to implement this EBP with fidelity. The RESET EI observation protocol offers one way to help close the research to practice gap by providing teachers with reliable evaluations of the quality with which they implement EI. Future research should investigate the use of the RESET EI observation protocol in practice contexts to ensure its successful adoption as a fidelity assessment.

References

Barnes, T. N., Cipriano, C., Flynn, L., Rivers, S. E., & Xu, W. (2019). Validating the Recognizing Excellence in Learning and Teaching (RELATE) Tool for special education classrooms. *The Journal of Experimental Education*, *87*(3), 415-429.

Bauer, M. S., Damschroder, L., Hagedorn, H., Smith, J., & Kilbourne, A. M. (2015). An introduction to implementation science for the non-specialist. *BMC psychology*, *3*(1), 32.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. Educational Assessment, 17(2-3), 62-87.

Bond, T. G., & Fox, C. M. (2007). *Fundamental measurement in the human sciences.* Chicago, IL: Institute for Objective Measurement.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation science*, *2*(1), 40.

Ciullo, S., Lembke, E. S., Carlisle, A., Newman Thomas, C., Goodwin, M., & Judd, L. (2016). Implementation of evidence-based literacy practices in middle school response to intervention: An observation study. *Learning Disability Quarterly, 39*(1), 44-57.

Cook, B. G., & Odom, S. L. (2013). Evidence-based practices and implementation science in special education. *Exceptional children*, *79*(2), 135-144.

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*, 267-297. doi:10.1002/pam.21818

Eckes, T. (2011). *Introduction to many-facet Rasch measurement*. Frankfurt, Germany: Peter Lang.

Englehard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5,* 171-191.

Gilmour, A. F., Fuchs, D., & Wehby, J. H. (2019). Are students with disabilities accessing the curriculum? A meta-analysis of the reading achievement gap between students with and without disabilities. *Exceptional Children, 85,* 329-346. doi:10.1177/0014402918795830

Harn, B., Parisi, D., & Stoolmiller, M. (2013). Balancing fidelity with flexibility and fit: What do we really know about fidelity of implementation in schools? *Exceptional Children, 79*(2), 181-193.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56-64.

Hill, H., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, *83*(2), 371-384.

Hughes, C. A., Morris, J. R., Therrien, W. J., & Benson, S. K. (2017). Explicit instruction: Historical contemporary contexts. *Learning Disabilities Research & Practice*, *32*(3), 140-148.

Justice, L. M., Mashburn, A. J., Hamre, B. K., & Pianta, R. C. (2008). Quality of language and literacy instruction in preschool classrooms serving at-risk pupils. *Early childhood research quarterly*, *23*(1), 51-68.

Kane, M. T. (2006) Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport, CT: Praeger.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Linacre, J. M. (2014). *A user's guide to FACETS Rasch-model computer programs*. Retrieved December, 18, 2018.

Linacre, J. M. (2019). Facets computer program for many-facet Rasch measurement, version 3.81.0. Beaverton, Oregon: *Winsteps.com*.

Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, *31*(1), 61-95.

Mantzicopoulos, P., French, B. F., Patrick, H., Watson, J. S., & Ahn, I. (2018). The stability of kindergarten teachers' effectiveness: a generalizability study comparing the framework for teaching and the classroom assessment scoring system. *Educational Assessment*, *23*(1), 24-46.

McKenna, J. W., Shin, M., & Ciullo, S. (2015). Evaluating reading and mathematics instruction for students with learning disabilities: A synthesis of observation research. *Learning Disability Quarterly, 38*(4), 195-207.

McLeskey, J., Barringer, M. D., Billingsley, B., Brownell, M., Jackson, D., Kennedy, M.,& Ziegler, D. (2017). High-leverage practices in special education. *Arlington, VA: Council for Exceptional Children & CEEDAR Center.*

McLeskey, J., Billingsley, B., Brownell, M. T., Maheady, L., & Lewis, T. J. (2019). What are High-Leverage Practices for special education teachers and why are they important?. *Remedial and Special Education*, *40*(6), 331-337.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-67.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American journal of evaluation*, *24*(3), 315-340.

Munter, C. (2014). Developing visions of high-quality mathematics instruction. *Journal for Research in Mathematics Education*, *45*(5), 584-635.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of applied measurement*, *4*(4), 386-422.

Odom, s. L. (2009). The tie that binds: Evidence-based practice, implementation science, and outcomes for children. *Topics in Early Childhood Special Education, 29,* 53-61. http://dx.doi.org/10.1177/0271121408329171

Odom, S. L., Hall, L. J., & Steinbrenner, J. R. (2020). Implementation science research and special education. *Exceptional Children, 86*(2), 117–119. https://doi.org/10.1177/0014402919889888

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78,* 33-84. http://dx.doi.org/10.3102/0034654307313793

Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*, 123-141. doi:10.17763/haer.82.1.v40p0833345w6384

Schulte, A. C., & Stevens, J. J. (2015). Once, sometimes, or always in special education: Mathematics growth and achievement gaps. *Exceptional Children, 81*, 370 –387. doi:10.1177/0014402914563695

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching project. *Education Finance and Policy*, *10*, 535-572. doi: 10.1162/EDFP_a_00173

Stockard, J., Wood, T. W., Coughlin, C., & Rasplica Khoury, C. (2018). The effectiveness of Direct Instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research, 88 (4),* 479-507.

Sullivan, K., Bell, N., Jones, D. H., Caverly, S., & Vaden-Kiernan, M. (2016). Implementation Work at Scale: An Examination of the Fidelity of Implementation Study of the Scale-Up Effectiveness Trial of Open Court Reading. *Society for Research on Educational Effectiveness*.

Swanson, E. A. (2008). Observing reading instruction for students with LD: A synthesis. *Learning Disability Quarterly, 31*, 1–19. doi:10.1177/0022219411402691

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, *102*(3628–3651. doi:10.1257/aer.102.7.3628

Vogel, S. P., & Engelhard Jr, G. (2011). Using Rasch measurement theory to examine two instructional approaches for teaching and learning of French grammar. *The Journal of Educational Research*, *104*(4), 267-282.

Wei, X., Blackorby, J., & Schiller, E. (2011). Growth in reading achievement of students with disabilities, ages 7 to 17. *Exceptional Children*, *78*, 89-106. doi:10.1177/001440291107800106

Figure 1

EI Rubric Items

	Explicit Instruction Rubric Item at Implemented Performance Level Descriptor
1	The goals of the lesson **are clearly** communicated to the students.
2	The stated goal(s) is/are **specific.**
3	The teacher **clearly** explains the relevance of the stated goal to the students.
4	Instruction is **completely aligned** to the stated or implied goal.
5	**All** of the examples or materials selected **are aligned** to the stated or implied goal.
6	Examples or materials selected **are aligned** to the instructional level of **most or all** of the students.
7	The teacher **effectively** reviews prior skills and/or engages background knowledge **before beginning** instruction.
8	The teacher **provides clear** demonstrations of proficient performance.
9	The teacher **provides an adequate number** of demonstrations given the nature and complexity of the skill or task.
10	The teacher uses language that is **clear, precise, and accurate** throughout the lesson.
11	Scaffolding is provided **when it is needed** to facilitate learning.
12	Complex skills or strategies **are broken down** into logical instructional units to address cognitive overload, processing demands, or working memory.
13	The teacher **systematically withdraws** support as the students move toward independent use of the skills.
14	Guided practice is **focused** on the application of skills or strategies related to the stated or implied goal.
15	The teacher **consistently prompts** students to apply skills or strategies throughout guided practice.

16	The teacher maintains an **appropriate** pace **throughout the lesson**.
17	The teacher **allows adequate** time for students to think or respond **throughout** the lesson.
18	The teacher **maintains focus** on the stated or implied goal **throughout** the lesson.
19	The teacher **provides frequent** opportunities for students to engage or respond during the lesson.
20	There are **structured and predictable** instructional routines throughout the lesson.
21	The teacher **monitors** students to ensure they remain engaged.
22	The teacher **consistently** checks for understanding **throughout the lesson**.
23	The teacher provides timely feedback **throughout the lesson**.
24	Feedback is specific and informative **throughout** the lesson.
25	The teacher **makes** adjustments to instruction **as needed** based on the student responses.

Figure 2

Variable map of the EI facets items, teachers, raters, and lessons

Measr	-Items	+Teacher	-Rater	-Lesson	Scale
4					(3)
3		11 45			
		38			
	I3				
2		12			
		1 30			
		14 34 37			
		19 44			---
		15 29 43 47			
1	I13	22 36 41 46 9			
		2 27 3 32 8			
	I25	6			
	I2	17 23 26	17		
	I7	48 7	11		
	I1 I24	25 5	12 4		
	I8 I9	10 16 18 21 28 31 40 42	1 2 5		
		35	16 19	3	
0	I12 I15 I16		18 6	2	2
	I11	13	20 8	1	
	I22	20	10 13 3		
	I10 I14 I17 I18 I23 I4		14 15		
	I20 I6	33 39			
			7 9		
	I21 I5	4			
-1	I19	24			

-2					(1)

Table 1

Validity Argument and Assumptions for RESET

1. Scoring assumptions
1.1 The scoring rule is appropriate.
1.2 Raters' understanding of the items are accurate and consistent with the developers' understanding.
1.3 Raters can consistently apply the scoring criteria.
1.4 Raters use the items without bias in that the same instructional behaviors and quality observed across different teachers would be scored similarly.
2. Generalizability assumption
2.1 Sufficient variance lies at the teacher level (not the rater or lesson level) to allow for reliable estimates of the quality of teachers' instruction.
3. Extrapolation assumptions
3.1 RESET consists of a set of distinct rubrics that detail the elements of evidence-based practices for students with high incidence disabilities. Performance across a set of items on an individual rubric represents the teacher's ability to implement the specific practice detailed in the rubric (trait interpretation).
3.2 Higher scores on a RESET rubric is positively related to student gains in the specific academic area (e.g. performance on decoding instruction is related to a student's reading growth)
3.3 Items accurately represent the evidence-based practices
4. Decision assumptions
4.1 Feedback to teachers based on RESET scores appropriately reflects key teacher strengths and weaknesses.
4.2 Conclusions reached using RESET are valid, in that the instrument performs at a minimum required level of reliability and accuracy.

Table 2

Participant Special Education Teacher Demographics

Demographic Variable	*N*
Gender	
Male	48
Female	0
Age	
20-29	10
30-39	14
40-49	11
50-59	9
60 or above	2
Missing	2
Ethnicity	
White, non-Hispanic	43
Asian	2
Latino	1
Two and more races	2
Education	
Bachelor degree	32
Master degree	14
Missing	2
Teaching Subjects	
Reading	33
Math	12
Reading and math	3
Total	48

Table 3

Item Measure Report from Many-Facet Rasch Measurement Analysis

Item Number	Fair Average	Difficulty (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
3	1.42	2.24	.08	1.56	1.63
13	1.93	.94	.07	1.16	1.16
25	1.93	.69	.07	1.13	1.13
2	2.09	.58	.07	1.19	1.17
7	2.11	.52	.07	1.26	1.25
1	2.18	.37	.07	1.16	1.15
24	2.18	.37	.07	.94	.99
8	2.25	.21	.07	.92	.98
9	2.25	.21	.07	1.00	1.04
12	2.33	.01	.07	1.04	1.06
16	2.34	.00	.07	.90	.92
15	2.34	-.01	.08	.92	.91
11	2.39	-.13	.08	.92	.96
22	2.42	-.23	.08	.90	.92
14	2.47	-.35	.08	.98	.96
10	2.47	-.36	.08	.88	.93
4	2.48	-.37	.08	.84	.89
23	2.49	-.41	.08	.88	.86
17	2.49	-.42	.08	.78	.82
18	2.49	-.42	.08	.86	.88
6	2.51	-.47	.08	1.06	1.11
20	2.52	-.49	.08	.96	.93
21	2.60	-.74	.08	.85	.90
5	2.60	-.74	.08	.87	.90
19	2.67	-1.01	.09	.86	.87
Mean (count =25)	2.32	.00	.08	.99	1.01
SD	.27	.67	.00	.17	.17

Note. Root mean square error (model) = .08; adjusted *SD* = .67; separation = 8.71; reliability = .99; fixed chi-square = 1703.3; df = 24; significance = .00.

Table 4

Teacher Measure Report from Many-Facet Rasch Measurement Analysis

Teacher Number	Fair Average	Ability (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
11	2.89	3.05	.20	1.24	1.18
45	2.89	3.03	.20	1.13	1.31
38	2.84	2.60	.18	1.24	1.11
12	2.73	2.02	.15	1.06	1.15
30	2.59	1.50	.12	1.12	1.02
1	2.58	1.47	.06	1.01	.97
34	2.56	1.39	.12	1.08	1.06
37	2.55	1.37	.13	1.13	1.22
14	2.54	1.35	.11	1.01	1.05
44	2.52	1.26	.13	1.00	1.03
19	2.50	1.21	.11	1.11	1.11
15	2.48	1.17	.11	.88	.88
29	2.45	1.09	.12	.99	1.02
43	2.45	1.08	.12	.91	.91
47	2.45	1.07	.12	1.22	1.23
46	2.44	1.04	.11	.99	.97
22	2.43	1.03	.12	.97	.92
9	2.42	1.00	.11	1.27	1.25
41	2.40	.96	.11	1.27	1.26
36	2.40	2.40	.11	.98	1.02
3	2.40	.93	.11	1.05	1.10
8	2.39	.92	.11	.94	.94
27	2.38	.90	.11	1.06	1.08
2	2.38	.89	.05	.96	.96
32	2.36	.84	.12	1.14	1.14
6	2.30	.69	.10	.99	.88
17	2.29	.66	.10	1.07	1.04
23	2.27	.62	.11	.82	.81
26	2.27	.61	.11	1.10	1.08
48	2.23	.53	.11	.69	.70
7	2.22	.49	.11	.92	.92
25	2.19	.44	.11	1.03	1.02
5	2.15	.34	.10	1.28	1.28
40	2.13	.29	.10	.88	.88
42	2.13	.29	.10	1.11	1.10
16	2.12	.27	.11	.93	.95
28	2.12	.26	.11	.99	1.06
31	2.10	.22	.11	.85	.84
18	2.09	.21	.11	.89	.89
10	2.09	.21	.10	.75	.74

21	2.09	.20	.11	1.03	1.09
35	2.05	.10	.10	.82	.80
13	1.93	-.16	.10	1.49	1.65
20	1.90	-.22	.11	1.11	1.09
39	1.80	-.45	.11	.76	.81
33	1.77	-.53	.11	.81	.81
4	1.70	-.70	.11	.66	.66
24	1.60	-.94	.12	.93	.94
Mean (count =48)	2.30	.78	.11	1.01	1.02
SD	.29	.82	.02	.17	.18

Note. Root mean square error (model) = .12; adjusted *SD* = .81; separation = 6.98; reliability = .98; fixed chi-square = 1905.0; df = 47; significance = .00.

Table 5

Rater Measure Report from Many-Facet Rasch Measurement Analysis

Rater Number	Severity	Model SE	Infit MNSQ	Outfit MNSQ
17	.62	.07	1.09	1.12
11	.47	.06	.79	.82
12	.41	.06	.77	.78
4	.37	.07	.74	.77
5	.30	.07	1.11	1.08
1	.26	.07	.90	.90
2	.25	.07	.98	1.03
19	.18	.07	1.06	1.05
16	.17	.07	.89	.92
6	.01	.07	.80	.81
18	-.04	.07	1.12	1.15
20	-.12	.07	.88	.91
8	-.15	.07	1.27	1.19
3	-.21	.07	1.06	1.05
10	-.28	.07	1.04	1.04
13	-.31	.07	1.06	1.11
15	-.31	.07	1.23	1.24
14	-.36	.07	1.19	1.21
7	-.59	.08	.88	.89
9	-.67	.07	1.23	.89
Mean (count =20)	.00	.07	1.00	1.01
SD	.36	.00	.16	.15

Note. Root mean square error (model) = .07; adjusted *SD* = .35; separation = 5.17; reliability = .96; fixed chi-square = 518.3; df = 19; significance = .00.

Table 6

Lesson Measure Report from Many-Facet Rasch Measurement Analysis

Lesson Number	Difficulty	Model SE	Infit MNSQ	Outfit MNSQ
3	.16	.03	1.00	1.00
2	-.02	.03	.96	.98
1	-.14	.03	1.03	1.05
Mean (count =3)	.00	.03	1.00	1.01
SD	.15	.00	.04	.04

Note. Root mean square error (model) = .03; adjusted *SD* = .15; separation = 5.59; reliability = .97; fixed chi-square = 69.2; df = 2; significance = .00.